Operational

Anonymisation and Data Protection

Guidance for staff and students

This guidance covers the anonymisation and/or pseudonymisation of personal data for research purposes. If you collect, store, process or transfer personal data for research purposes you need to be aware of these techniques, why they are required, where they are applicable and who to contact if you have any questions or concerns relating to them.

How can I share research data without breaking the law?

Please Note:

- Information presented here should be read alongside LSE's Data Protection Policy [PDF]. Address questions on data protection to the school's Data Protection Officer.
- Personal data that directly, or indirectly in combination with other data sources identifies a living individual and is governed in the UK by the Data Protection Act (2018) and the General Data Protection Regulation (GDPR) [PDF] [Data Protection legislation].
- Data Protection legislation does not apply when anonymised data cannot be linked to a living individual. This means any identifying data would need to be removed before the dataset could be considered anonymised.

Data Protection legislation governs the processing of data or information for living individuals in the UK. The legislation requires data be handled in a way that is fair, proportional, secure, and justified when obtaining, using, holding and sharing personal information. LSE has guidance [PDF] on how to meet these principles in the context of research.

There are two additional things to remember about the legislation in relation to research. First, it only applies to personal or special categories personal data, not necessarily all data gathered from a participant. Second, the legislation contains exemptions for specified purpose and retention of personal data when processed for research.

Anonymised data that cannot be linked to a living individual is not subject to Data Protection legislation, though there may still be ethical reasons for protecting this information.

What counts as “anonymised” is measured by a “likely reasonably” test. The UK’s Information Commissioner’s Office states: "Anonymisation is the process of turning data into a form which does not identify individuals and where identification is not likely to take place." This means that if, on the balance of probabilities, third parties cross-referencing “anonymised” data with information or knowledge already available to the public cannot identify individuals then data is not personal and not subject to the legislation.

For a very thorough checklist on anonymising and/or pseudonymising data, the US Bureau of the Census provides the following checklist: https://www.census.gov/srd/sdc/drbchecklist51313.docx.

How can I anonymise research data to protect participants?

Please Note:

- Anonymisation refers to direct identifiers and indirect ones which in combination can identify an individual.
- Plan and apply anonymisation early in your research and log changes so it is clear what is anonymised.
- Anonymisation tools are available.
- Consider alternative access options like controlled access environments and restrictive licences where sharing anonymised data is problematic.

Usually anonymisation applies to direct and indirect identifiers. Direct identifiers like name, address, or telephone numbers specify an individual. Indirect identifiers when pieced together could also reveal an individual by, for example, cross-referencing occupation, employer, and location.

If data requires anonymising, it’s critical to think early on about how you are going to construct and implement a strategy to protect the identity of participants. Planning anonymisation before undertaking data collection produces both better informed consent and requires a less resource intensive process when doing data anonymisation.

Given the strength of the GDPR, it is worth questioning what data you plan to collect and why

Knowing what data you wish to collect will help guide an anonymisation strategy consistent across your data set and produce ethically responsible reusable data that does not contravene data protection laws. For example, administrative data like names and addresses may not have research value but constitute personal and sensitive information. Do they need to be collected, if so, can they be separated from the research data set and deleted early in the research process?

Remove direct identifiers or use meaningful pseudonyms and replacements for identifiers. Ideally, replacements should be expressive in the sense of preserving the character of the identifier while concealing the identity. For example, instead of “Birmingham” use “Major British metropolitan area” and instead of “Scott” use “Trevor”. This is preferable to replacing identifiers with “City” or “Name” or, worst of all, “deleted”.

If using pseudonyms is unworkable, could you apply restrictions on upper and lower ranges of

variables? Can you remove a variable without it compromising the re-use value of the data (in which case, ask if you should you even measure that variable)? Could you apply low-level aggregation of data, like moving to a larger spatial unit or transforming age from a continuous variable into a discrete categorical one? Can dates, times, or measurements be rounded? In any case, it is best practice to create a log of anonymisation undertaken and to flag anonymised identifiers so it is clear that something is anonymised.

There are a number of Open Source tools developed to help researchers anonymise research data.

- Cornell Anonymization Toolkit (quantitative)
- sdcMicro, (quantitative) software for R and paper [PDF] on using the software
- IQDA Qualitative Data Anonymiser and instructions [PDF] (note: its developer is still testing this tool).

Finally, can data be shared in a restricted environment? Most standard archive and data re-use agreements have a clause prohibiting third party attempts to either identify or re-contact participants. In other cases, controlled access environments and applying approved researcher status may be a way to responsibly share research data to a limited extent.

Of course, the principle of informed consent allows participants to waive their right to anonymity should they wish, and if in the researcher's judgment, no harm will result or no other legal reason exists to prevent waiving anonymity. In the case of oral history or elite interviews, tied to the participant's identity are their memories, perceptions and experiences. Consequently, data in these approaches is not anonymised even if it is subject to tighter access conditions or an embargo period.

The Data Protection Act (2018), General Data Protection Regulation (GDPR) and research

Please Note:

- The GDPR is a European Economic Area wide regulation on data protection and will be automatically brought into UK law at the point that Brexit occurs.
- The regulation is supplemented by the UK Data Protection Act 2018.
- It expands the definition of personal data, includes tougher penalties for data protection breaches, and stronger requirements for obtaining informed consent than previous legislation.

GDPR is a European Union directive governing the protection of personal data inside and outside the EU. It is supplemented in UK law by the Data Protection Act (2018).

For researchers, GDPR provisions do not substantively differ from previous UK law. It protects exemptions for research data on reuse but tightens the principle of informed consent, expands the definition of personal data, and contains stronger penalties for data protection breaches.

Special categories personal data includes an individual's racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, health, sex life or sexual orientation. The law now includes genetic and biometric data as well as online identifiers. Data on an individual's criminal convictions is also included but is subject to even tighter controls.

Informed consent is a strong theme throughout the regulation. While consent isn't the only grounds for processing personal data, from a research perspective lawful processing of personal data is almost always based on consent (Article 6, 1).

The UK Data Protection Act defines consent as being lawful from the age of 13 although the regulation allows other EU member states to vary that age to as high as 16 (Article 8, 1). There is a requirement to be able to demonstrate consent has been given (Article 7, 1) and that subjects have the right to withdraw consent (Article 7, 3). Informed means the subject is aware of who is collecting the data and why. The Regulation states: "Consent should not be regarded as freely given if the data subject has no genuine or free choice or is unable to refuse or withdraw consent without detriment." (recital 42)

The regulation does allow for exemptions on revealing personal data where explicit consent is allowed by law and has been given by the subject (Article 9,1).

The regulation reinforces the requirement of those collecting data to safeguard personal data. Safeguards include technical and organisational barriers to access, like an encryption, authentication requirements and user licences, or applying anonymisation or pseudonymisation that would "no longer permit the identification of data subjects" (Article 89, 1). Significant fines can be imposed for breaches where data could have been better protected and was not.

Personal data may only be transferred outside of the European Economic Area where you either have explicit and informed consent to do so or where the Commission has decided a country, territory, sector, or international organisation ensures an adequate level of protection and redress (Chapter 5)

The publicised "right to be forgotten" is in the regulation (Article 17). However, that right does not apply where personal data is necessary for "archiving purposes in the public interest or for scientific statistical and historical purposes" nor can it apply if the data is anonymised or pseudonymised to prevent identification of living individuals.

The regulation reinforces the requirement of those collecting data to safeguard personal data. Safeguards include technical and organisational barriers to access, like an encryption, authentication requirements and user licences, or applying anonymisation or pseudonymisation that would "no longer permit the identification of data subjects" (Article 89, 1). Significant fines can be imposed for breaches where data could have been better protected and was not.

If you're planning a research project it's important to factor in the effects of the legislation in your consent and data protection planning.

Data Protection Impact Assessment

You may be asked by a data supplier, public body or private organisation to provide a Data Protection Impact Assessment (DPIA), also known as a Privacy Impact Assessment (PIA). A DPIA is a process used to identify any potential privacy risks and describe actions to address them, so even if you are not required to complete one, it is worth looking through the screening questions to determine if you should anyway. LSE has a template [.docx] to help you write a Data Protection Impact Assessment with support from IMT Information Security and Information Rights teams.

Further reading

CESSDA ERIC (2018) Processing personal data

Finnish Social Science Data Archive Anonymization and Identifiers

Irish Qualitative Data Archive (2008) Anonymisation Guidelines [PDF]

Information Commissioners Office (2012)Anonymisation: Managing Data Protection [PDF]

Risk Code of Practice JISC Legal (2014) Data Protection and Research Data: Questions and Answers

UK Anonymisation Network (2016) Anonymisation Decision-Making Framework

<NEW REFERENCES FROM PD to add?>

Managing and Sharing Research Data: A Guide to Good Practice (2019) Corti, L et. al. (October 2019) 368pp. SAGE Publications Ltd

https://www.jisc.ac.uk/guides/cloud-computing-innovation-gdpr-and-you

https://ico.org.uk/media/1061/anonymisation-code.pdf

https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2019/08/data-minimisation-and-privacy-preserving-techniques-in-ai-systems/

Questions and Concerns

If you are have any questions or concerns relating the use of personal data for research purposes please contact the School's Data Protection Officer Rachael Maguire (glpd.info.rights@lse.ac.uk).

'If you have any questions about preparing data for publication or archiving please contact the Research Data Librarian (datalibrary@lse.ac.uk).

Review schedule

Review interval	Next review due by	Next review start
3 years	1/1/2023	1/1/2023

Version history

Version	Date	Approved by	Notes
0.1	03/11/2019	Draft for review only	
1.0	01/06/2020	IGMB	

Links

Reference	Link

Contacts

Position	Name	Email	Notes
Data Protection Officer	Rachael Maguire	r.e.maguire@lse.ac.uk	

Communications and Training

Will this document be publicised through Internal Communications?	**Yes/ No**
Will training needs arise from this policy	**Yes/ No**
If Yes, please give details	